A words-of-interest model of sketch representation for image retrieval

Conference or Workshop Item

A words-of-interest model of sketch representation for image retrieval

Xi Luo[#1], Wen-Jin Guo[#2], Yong-Jin Liu[#3], Cui-Xia Ma[*4], Da-Wei Song[^5]

[#]*Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[1]lxljs81@163.com
[2]wenjing103@gmail.com
[3]liuyongjin@tsinghua.edu.cn

[*]*Intelligence Engineering Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China*
[4]cuixia.ma@gmail.com

[^]*School of Computing, The Robert Gordon University, United Kingdom*
[5]d.song@rgu.ac.uk

***Abstract*— In this paper we propose a method for sketch-based image retrieval. Sketch is a magical medium which is capable of conveying semantic messages for user. It's in accordance with user's cognitive psychology to retrieve images with sketch. In order to narrow down the semantic gap between the user and the images in database, we preprocess all the images into sketches by the coherent line drawing algorithm. During the process of sketches extraction, saliency maps are used to filter out the redundant background information, while preserve the important semantic information. We use a variant of Words-of-Interest model to retrieve relevant images for the user according to the query. Words-of-Interest (WoI) model is based on Bag-of-visual Words (BoW) model, which has been proven successfully for information retrieval. Bag-of-Words ignores the spatial relationships among visual words, which are important for sketch representation. Our method takes advantage of the spatial information of the query to select words of interest. Experimental results demonstrate that our sketch-based retrieval method achieves a good tradeoff between retrieval accuracy and semantic representation of users' query.**

***Keywords*— Image retrieval, Sketch representation, Bag-of-visual Words model, Words-of-Interest model, Markov chain model**

I. INTRODUCTION

With the fast advances of digital imaging equipments and computer network technology, tons of images are accessible on the Internet. With the help of image processing software, like Photoshop, designer can make an amazing picture by compositing images together. This requires the designer to provide images with specific objects or content for composition, which can be figured out by solving an image retrieval problem. Image retrieval is one of key technologies in improving people's life quality in terms of reuse of information and knowledge. Content-based image retrieval has attracted much attention in the past few decades [19].

Text retrieval is one of the traditional and common methods of image retrieval with a good semantic understanding. In this method, the system needs to provide annotation for every image in databases. Manually image annotation is time-consuming, laborious and expensive, since massive semantic information is needed to discriminate different images. Automatic image annotation methods could tag captioning or keywords to a digital image automatically, which makes text retrieval system more possible. Although queries can be more naturally specified by the user in text retrieval system than other forms of retrieval system, the potential users' language difference and inconsistent naming issues, make the text retrieval system unpractical in certain conditions.

Different from text retrieval, image retrieval with image as input doesn't need to annotate the images in databases. To shorten the on-line retrieval time, images are preprocessed into feature vectors based on certain feature extraction algorithms. Since the feature extraction algorithms have nothing to do with the users, the feature vector is always identical for the same image. So it's not necessary for common users to know the details of the feature extraction algorithms and retrieval methods. Users only need to provide a sample image, and then the retrieval system will retrieve the relevant images in the database. But this method will not work in the absence of sample images.

Sketch-based image retrieval is a good solution to the problem of sample images' absence. Sketches present natural description facilitating the image semantics, as it is consistent with human cognition [20]. So sketches could be used to represent an image. Sketches have also been proved to represent videos successfully [15][16]. So sketch is a very good medium to preserve information. With some shape

modelling technology, 3D shape could even be recovered from sketches [24]. In this paper, we propose a method for sketch-based image retrieval. All the images in the database are preprocessed into sketches by a coherent line drawing algorithm [9]. During the process of sketches extraction, saliency maps are used to filter out the redundant background information, while preserve the important semantic information. Our method works on the sketch level, using a words-of-interest (WoI) model. Experiment results show that our image retrieval method achieves good performance.

The remainder of this paper is organized as follows. Section II summarizes the related work. Section III presents the methods of sketch extraction from images and feature extraction from sketches. Section IV provides the experiment results of our system. Section V provides the concluding remarks.

II. Related Work

The QVE (query by visual example) system proposed by Kato et al.[1] is the first method which allows user to draw an outline sketch of the general image composition as input. The user's sketch is matched to the abstract images, with a map of effective edge points extracted from the original full color images. Local similarity between the corresponding local blocks of the user's sketch and the abstract image, is calculated as the maximum local-correlations by shifting the two blocks. The global similarity is a sum of the local similarities. IBM's QBIC system [2] is a modified version of this approach. But this approach doesn't allow indexing and the used feature vector doesn't have the property of rotation invariance. The method proposed by Bimbo et al. [3] retrieves visual images by elastic matching of user sketches. This method is expensive and also does not have rotation invariant property.

Agouris et al. [4] present a sketch-based retrieval system for retrieving digital images from topographic databases. Their system retrieves images with metadata information and a sketch is ued as input. The metadata information could be about the general properties of the image itself, or include additional information such as time and location of image acquisition, scale, resolution, etc. The metadata information is used to narrow down the potential matches before query by sketch begins. Chalechale et al. [5] proposed an angular-radial decomposition algorithm for sketch-based image retrieval. It's able to measure the similarity between the full color database images and the sketched query. This method derives abstract images based on edges of the database image and thinned outline of the sketched query. Features are extracted by partitioning the abstract images angular-radial sectors. Fourier transformation is used to get rotation invariance. But the effectiveness of this algorithm is not so promising when compared to state-of-the-art retrieval algorithms.

The method proposed by Anelli et al. [6] aims at sketch-based image retrieval in complex scenes. They propose a variant of Generalized Hough transform to solve two main problems of sketch-based image retrieval systems: (1) an inexact matching problem and (2) the selection of the image features when comparing with the user's sketch. This method is suitable for matching a sketch with edge images extracted from the database images. But the time complexity for similarity calculation makes this method not suitable for real-time image retrieval in large-scale databases [7]. Saavedra et al. [8] proposed a method based on a histogram of edge local orientations (HELO) for sketch-based image retrieval. This method is invariant to scale, translation and rotation. But HELO descriptor may fail to represent simple models discriminatively, since HELO descriptor is a global representation.

III. Sketch and Feature Extraction

A. Sketch Extraction

Sketch is a natural and effective way for user to express their ideas [23]. To narrow down the semantic gap between the user and the images in database, we preprocess all the images in the database into sketches by the line drawing algorithm [9]. Sketch, extracted from the original image, throws away some redundant and useless information of the original image, and keeps the important semantic outline information. But the resulting sketch by only using the line drawing algorithm [9] tends to also extract the outline of the background. The mixture of the background outline and the foreground object outline will lead to irrelevant retrieval results. So in our method, before the process of sketch extraction, the algorithm proposed by Cheng et al. [10] is used to extract a saliency map for every image in the database. Saliency maps are used to filter out the redundant background information, while preserving the important foreground objects information. Then, the line drawing algorithm [9] is used in salient region to extract sketches from the filtered images. Some results of the sketch extraction are showed in Fig. 1.

B. Feature Extraction

Our algorithm is based on Words-of-Interest selection model, proposed by Wang et al. [11]. Words-of-Interest model takes advantage of the spatial information of the query to select words of interest. In our paper, we use shape profile to extract the local feature descriptors. Then a visual vocabulary is generated from the extracted features of training images, using a k-means clustering method. Finally, a Markov chain model is adopted to select words-of-interest from the visual vocabulary according to the query.

1) *Shape Profile:* In our method, we randomly sample 500 points in the sketch image. With every sample point as a reference point, we extract features based on the distribution of sketch lines. Of the same object, different users will draw sketches of different styles. Even the two sketches of a same object, drawn by the same person but at different time, are probably quite different. Due to the uncertainty of the strokes' directions and positions, it's not reasonable to force the sample points to lie on sketch lines or lie in the empty areas. So we sample points both on sketch lines and in empty areas.

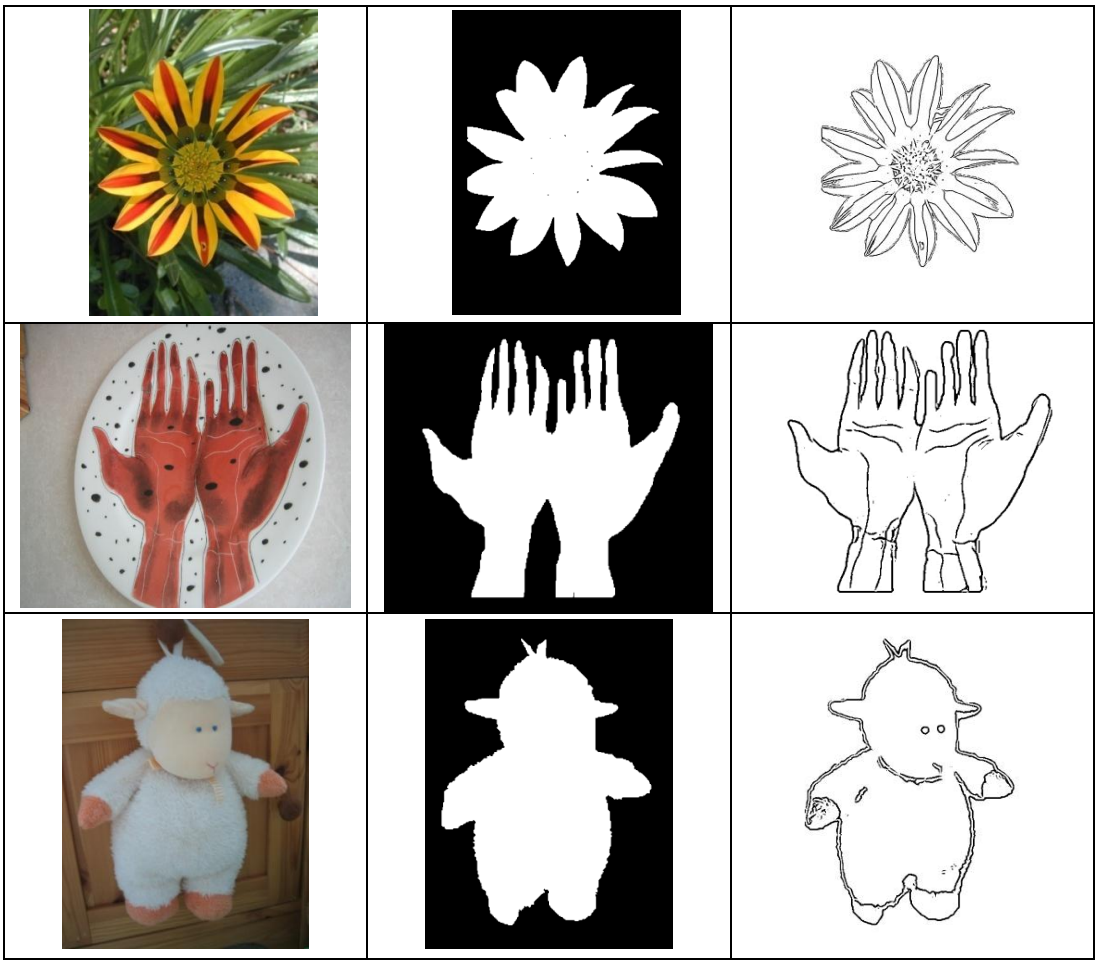


Fig. 1 Sketch extraction results. The left column: the original images. The middle column: the saliency maps. The right column: the sketches of the corresponding original images.

Instead of directly sampling points in the sketch image, we randomly sample points in the salient area of the saliency map. The salient area is usually not a regular region. To make the random sampling method simple and effective, we calculate a minimum bounding rectangle of the salient area. We randomly sample in the rectangle area. The rectangle area is considered as C-space formed from a Cartesian product, $S = X \times Y, X \in [0,1], Y \in [0,1]$. Two uniform random samples x and y are calculated from X and Y respectively, so (X, Y) is a uniform random sample for S. To restrict the sample points lie in the salient area, we discard the sample points which lie in the rectangle area but out of the salient area. Fig. 2 (a) shows the distribution of sample points in the salient area. Since the sketch image is extracted with the saliency map as a filter, the sample points' positions in the saliency map can be transferred to the sketch image without any change. But for the user's sketch, the situation is different since there is no corresponding saliency map. We calculate a minimum bounding rectangle, surrounding all the sketch lines. Then points are randomly sampled in the rectangle area. Or user could specify the salient area of the sketch.

The feature descriptor is defined as the distribution of sketch lines in the fixed local radius range of a sample point, as showed in Fig. 2(b). The local radius is equally divided into 20 bins. Every pixel in the sketch lines, which is in the fixed range of a sample point, makes a contribution to the corresponding bin. So the formula of the feature descriptor F is defined as follows:

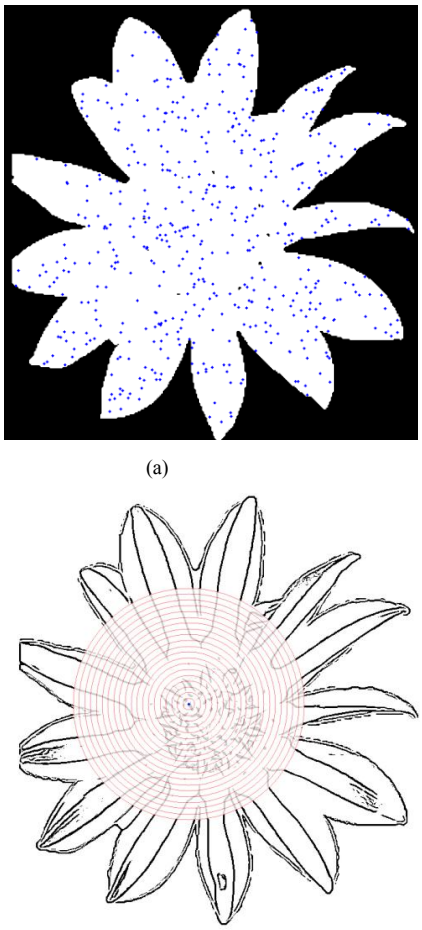
$$F(k) = \{\# p_i \in bin(k), k \in 1,2,...,20\} \quad (1)$$

Where k is the index of bin, p_i is a pixel on sketch lines, and $F(k)$ calculates the number of p_i which falls into bin k. $F(k)$ is normalized with the sum of pixels on sketch lines in the sketch image. The similarity of two feature vectors is defined as follows:

$$Sim = \frac{F_1 F_2}{\|F_1\| \|F_2\|} \quad (2)$$

2) *Words-of-Interest Model:* A set of features are extracted from images using shape profile as feature descriptor. K-means clustering is used to build the visual vocabulary based on a bag-of-words (BoW) model [21][22].

For a given query, some visual words in the visual vocabulary are more important than the other visual words. But BoW assigns all visual words with equal importance. The algorithm, proposed by Wang et al. [11], extracts words-of-interest (WoI) from BoW according to the query.



(a)

(b)

Fig. 2 The sampling result and one histogram used to calculate the distribution of the local sketch line pixels. (a) 500 sample points in the salient area. (b) A histogram of 20 radial bins centered at a sample point.

Inspired by the algorithm of Wang et al. [11], we use a Markov chain model to select WoI according to the query. In our approach, the visual vocabulary is considered as a finite state space of the Markov chain model. Visual words with higher probability to occur in the query are selected as WoI. Suppose the query is represented as a weighted vector of visual words: $w_q = \{N_i W_i\}, F_q = \{N_i\}, i = 1,...,N_w$, N_i is the term frequency of visual word W_i, N_w is the size of the visual vocabulary. The spatial proximity of two visual words is defined as the average similarity of visual words' instances:

$$S_{ij} = \frac{1}{N_i N_j} \sum_{m=1}^{N_i} \sum_{n=1}^{N_j} Sim_{m,n} \qquad (3)$$

Where N_i and N_j are the number of instances of visual words w_i and w_j respectively, $Sim_{m,n}$ is the inverse of the Euclidean pixel distance of the m^{th} instance of w_i and the n^{th} instance of w_j. The higher spatial proximity two visual words have, the more possible a vidual word would transfer to the other one. We define the visual word transfer probability matrix as :

$$P = [P_{ij}] = [S_{ij} / \sum_{j=1}^{N_w} S_{ij}] \qquad (4)$$

The conditional probability that visual word w_i occurs in the query is N_i / N_q, N_q is the number of features in the query. The initial state distribution of the Markov chain model is defined as

$$\pi(0) = \{N_i / N_q, i = 1,...,N_w\} \qquad (5)$$

The limit state distribution π^*, which indicates the visual words' final probability of occurrence, is calculated by the following formula:

$$\pi^* = \lim_{n\to\infty} \pi(n) = \lim_{n\to\infty} \pi(0) P^n \qquad (6)$$

According to π^*, the visual words are sorted with the probability of occurrence from high to low principle. The visual words on the top are selected as WoI. The distance between two sketches p and q based on WoI is defined as:

$$D = \alpha |F_p - F_q| + \beta |F_p' - F_q'| \qquad (7)$$

Where F_p and F_q are the term frequency of non-WoI of sketch p and q respectively, F_p' and F_q' are the term frequency of WoI of sketch p and q respectively, α and β $(\beta > \alpha)$ are the weighting factors.

Both a sketch and a document carry information, and only the representation forms are different. A sketch is represented as visual words, so we can consider a sketch as a document consisting of visual words. A Markov chain used to represent a document has been proved to be ergodic [13] [14]. So we have good reason to assume that the Markov chain model in our method is also ergodic. The strict proof of this assumption is left in the future work. An ergodic chain has a property that a stationary distribution exists and it equals the limit distribution. So the limit distribution π^* is irrespective of the initial state and represents a desirable distribution.

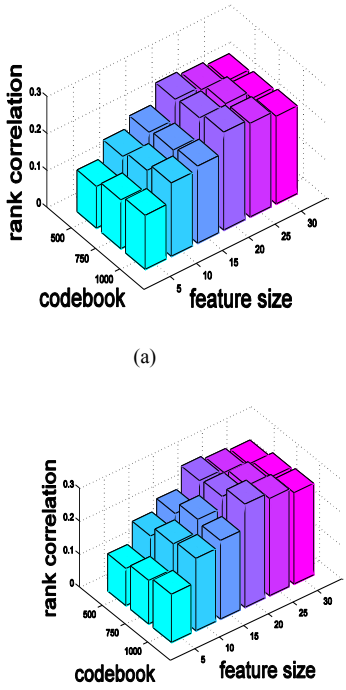
IV. Experiment Results

In this experiment, we evaluate the performance of our method using two public available benchmarks of image databases. One is a sketch-based benchmark database [12]. This database has two datasets, a benchmark dataset and a distractor image dataset. The benchmark dataset contains 31 benchmark sketches and 40 corresponding images for each sketch. The benchmark sketches are used as queries to rank

the corresponding images. The benchmark dataset and the distractor image dataset are joined together as an evaluation set. This database contains the rankings of the images, collected from the user study. The second database is the object retrieval database [17] from University of Kentucky. This database contains 10,200 images, with 2550 sets of four images of the same object. The metric of "top-4-score" is used to evaluate and compare the performance of our method with several classic methods [12][17][18].

A. *Results on Benchmark Database*

To measure the influence of different parameter values for this method, we set different values for the visual vocabulary size and local radius of shape profile. Different values for the sizes of a visual vocabulary are 500, 750, 1000. Different values for the local radii are 5, 10, 15, 20, 25, 30 (percentage of the diagonal of the training images' salient area). We use a set of 20,000 training images for visual vocabulary generation. The training images are randomly chosen from the distractor image dataset, and those images are excluded from the evaluation set. 500 features are extracted from every training image. K-means clustering is used to build the visual vocabulary of different sizes.



(a)

(b)

Fig. 3 Evaluations for BoW (a) and our method using WoI model (b)

Our method is used to rank the images based on the benchmark sketches. The correlation between the user's rankings and the objective rankings measures the performance of our method. To measure the influence of WoI model, we also measure the performance of our method replacing WoI model with BoW model. In Fig. 3 (a), it shows the rank correlation coefficient of the BoW model. In Fig. 3 (b), it shows the rank correlation coefficient of our method, using WoI model. So our method with WoI model has higher rank correlation coefficient than that using BoW model. The maximum correlation of our method is 0.313 for a vocabulary size of 1000 visual words and 20% of the salient area diagonal, while the maximum correlation coefficient in [12] is 0.277.

B. *Results on Object Retrieval Database*

From the object retrieval database, we randomly select 100 sets of images as a test dataset, and use the remaining sets of images as a training dataset. Features are extracted from the training dataset, with the local radius of shape profile as 20% of the salient area diagonal. A visual vocabulary of size 1,000 is generated from the training dataset. Every image in the test dataset is used as query to retrieve four top images. The number of true positives in the four top images is the score for this query. The average number of scores for all the queries is calculated to measure the performance of our method. Our method achieves a competitive retrieval performance with an average score of 3.12, which is a little bit lower than the average score of 3.60 in [18] but higher than the average score of 3.1 in [17]. Since sketches are generally well-known to be semantically meaningful for the user input, our presented method achieves a good balance between retrieval accuracy and semantic representation of users' query.

V. Conclusions

In this paper, we propose a method to retrieve images based on sketches. It's based on WoI model. We adopt a Markov chain model to incorporate the spatial information of the visual words to extract words of interest. The experiment results show that our method using WoI model improves the performance of that using BoW model.

Acknowledgment

The authors thank the reviewers whose comments help improve this paper very much. This work is partially supported by Projects of International Cooperation and Exchanges NSFC (Project Number 61111130210), the Natural Science Foundation of China (Project Number 60970099) and Tsinghua University Initiative Scientific Research Program (Project Number 20101081863).

References

[1] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A Sketch Retrieval Method for Full Color Image Database-Query by visual example," *Proc. of. Int. Conf. on Pattern Recognition*, pp.530-533, 1992.

[2] W. Niblack, R. Barber, W. Equitz, M.Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, " The QBIC project: querying images by content using color, texture, and shape," In *Proc. Of Spie*, USA, 1993, vol. 1908, pp.173-187.

[3] A. Del Bimbo, P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 2, 1997, pp. 121-132.

[4] P. Agouris, J. Carswell, A. Stefanidis, "Sketch-Based Image Queries in Topographic Databases," *Journal of Visual Communication and Image Representation*, 1999, 10:1-16.

[5] A. Chalechale,G. Naghdy, A. Mertins, "Sketch-Based image matching using angular partitioning," *IEEE Trans. on Systems, Man and Cybernetics*, Part A:Systems and Humans 35(1), 2005, pp.28-41.

[6] M. Anelli, L. Cinque, and E. Sangineto, "Deformation Tolerant Generalized Hough Transform for Sketch-Based Image Retrieval in Complex Scenes," *Image and Vision Computing*, vol. 25, no. 11, pp. 1802-1813, 2007.

[7] J. Revaud, G. Lavoue, A. Baskurt, "Improving Zernike Moments Comparison for Optimal Similarity and Rotation Angle Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31.no.4, 2009, pp. 627-636.

[8] J. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," *Proceedings of the 32nd DAGM conference on Pattern recognition*, 2010, pp.432-441.

[9] Kang H., Lee S., Chui C. K. "Coherent line drawing," In *Proc. Non-photorealistic Animation and Rendering* (2007), pp. 43–50.

[10] Ming-ming Cheng, Guo-Xin Zhang, Niloy J.Mitra, "Global Contrast based Salient Region Detection," *IEEE CVPR*, p. 409-416, Colorado Springs, USA, June 21-23, 2011.

[11] Lei Wang, Da-Wei Song and Eyad Elyan, "Video Retrieval based on Words-of-Interest Selection," *The 33rd European Conference on Information Retrieval (ECIR'2011)*, pp. 687-690. 19-21 April 2011.

[12] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur and Marc Alexa, "Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors," *IEEE Transactions on Visualization and Computer Graphics*, 2011, pp. 1-14

[13] Hoenkamp, E., Da-Wei Song, "The document as ergodic markov chain," In *Proceedings of ACM/SIGIR'2004*, Poster, pp. 496-497.

[14] Hoenkamp, E., Bruza, P.D., Da-Wei Song, Huang, Q., "An effective approach to verbose queries using a limited dependencies language model," *The 2nd International Conference on Theory of Information Retrieval (ICTIR'2009)*, LNCS 5766, pp. 116-127. 10-12 September 2009, Cambridge, UK.

[15] J. P. Collomosse, G. McNeill, L. Watts, "Free-hand sketch grouping for video retrieval," *Intl. Conf on Pattern Recognition (ICPR)* (Dec. 2008), pp. 1-4.

[16] J. P. Collomosse, G. McNeill, Y. Qian, "Storyboard sketches for content based video retrieval," *Intl. Conf on Computer Vision (ICCV)* (Sep. 2009), pp.245-252.

[17] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161-2168, June 2006.

[18] Zhong Wu, Qifa Ke, JianSun, Heung-Yeung Shum, "A Multi-Sample, Multi-Tree Approach to Bag-of-Words Image Representation for Image Retrieval," *ICCV09*.

[19] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp.1349-1380, 2000.

[20] V. Goel. *Sketches of Thought*, Cambridge, Mass.: MIT Press, 1995.

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," In *7th Int. WWW Conference*, 1998.

[22] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," *Intl. Conf on Computer Vision (ICCV)* (Sep. 2003).

[23] Cui-Xia Ma, Yong-Jin Liu, Hai-Yan Yang, Dong-Xing Teng, Hong-An Wang, Guo-Zhong Dai, "KnitSketch: A Sketch Pad for Conceptual Design of 2D Garment Patterns," *IEEE Transactions on Automation Science and Engineering*, Vol. 8, No. 2, pp. 431-437, 2011.

[24] Yong-Jin Liu, Kai Tang, Ajay Joneja, "Sketch-based free-form shape modelling with a fast and stable numerical engine," *Computers & Graphics*, Vol.29, No.5, pp.778-793, 2005.